Structural organization of the barley D-hordein locus in comparison with its orthologous regions of wheat genomes

Yong Qiang Gu, Olin D. Anderson, Cynthia F. Londeorë, Xiuying Kong, Ravindra N. Chibbar, and Gerard R. Lazo

Abstract: D hordein, a prolamin storage protein of barley endosperms, is highly homologous to the high molecular weight (HWM) glutenin subunits, which are the major determinants of bread-making quality in wheat flour. In hexaploid wheat (AABBDD), each genome contains two paralogous copies of HMW-glutenin genes that encode the x- and y-type HMW-glutenin subunits. Previously, we reported the sequence analysis of a 102-kb genomic region that contains the HMW-glutenin locus of the D genome from *Aegilops tauschii*, the donor of the D genome of hexaploid wheat. Here, we present the sequence analysis of a 120-kb D-hordein region of the barley genome, a more distantly related member of the Triticeae grass tribe. Comparative sequence analysis revealed that gene content and order are generally conserved. Genes included in both of these orthologous regions are arranged in the following order: a Xa21-like receptor kinase, an endosperm globulin, an HMW prolamin, and a serine (threonine) protein kinase. However, in the wheat D genome, a region containing both the globulin and HMW-glutenin gene was duplicated, indicating that this duplication event occurred after the separation of the wheat and barley genomes. The intergenic regions are divergent with regard to the sequence and structural organization. It was found that different types of retroelements are responsible for the intergenic structure divergence in the wheat and barley genomes. In the barley region, we identified 16 long terminal repeat (LTR) retrotransposons in three distinct nested clusters. These retroelements account for 63% of the contig sequence. In addition, barley D hordein was compared with wheat HMW glutenins in terms of cysteine residue conservation and repeat domain organization.

Key words: HMW glutenin, evolution, retrotransposon, comparative genomics.

Résumé : La D hordéine, une protéine de réserve de type prolamine dans l'albumen de l'orge, est très homologue aux sous-unités de grande taille des gluténines (HMW-gluténines), lesquelles déterminent en grande partie la qualité boulangère de la farine de blé. Chez le blé hexaploïde (AABBDD), chaque génome contient deux copies paralogues des gènes codant pour les HMW-gluténines, l'une codant pour les sous-unités de type x et l'autre pour les sous-unités de type y (Shewry et al., 1995). Antérieurement, les auteurs avaient analysé une région génomique de 102 kb contenant le locus d'une HMW-gluténine du génome D provenant de l'*Aegilops tauschii*, l'espèce ayant contribué le génome D chez le blé hexaploïde. Dans le présent travail, les auteurs ont analysé la séquence d'un segment génomique de 120 kb contenant le locus de la D hordéine chez l'orge, une espèce plus distante au sein des hordées. Une comparaison des séquences a révélé que le contenu et l'ordre géniques sont généralement conservés. Au sein de ces régions orthologues, on retrouve dans l'ordre des gènes codant pour une récepteur kinase de type Xa21, une globuline de l'albumen, une prolamine de grande taille et une sérine (thréonine) protéine kinase. Cependant, dans le génome D du blé, une région contenant les gènes codant pour la globuline et la HMW-gluténine était dupliquée, ce qui suggère que cette duplication a eu lieu après la séparation des génomes du blé et de l'orge. Les régions intergéniques étaient différenciées quat à leur séquence et leur organisation structurale. Il a été noté que différents types de rétroéléments étaient responsables de la divergence structurale observée dans les régions intergéniques entre le blé et l'orge. Dans le génome de l'orge, les auteurs ont identifié seize rétrotransposons à LTR formant trois groupes distincts. Ces rétroéléments totalisaient 63 %

Received 23 April 2003. Accepted 10 July 2003. Published on the NRC Research Press Web site at http://genome.nrc.ca on 30 October 2003.

Corresponding Editor: G.J. Scoles.

Y.Q. Gu,[1] O.D. Anderson, C.F. Londeorë, and G.R. Lazo. United States Department of Agriculture, Agricultural Research Service, Western Regional Research Center, 800 Buchanan Street, Albany, CA 94710, U.S.A.
X. Kong. Institute of Crop Germplasm Resources, Chinese Academy of Agricultural Sciences, Beijing 100081, China; and United States Department of Agriculture, Agricultural Research Service, Western Regional Research Center, 800 Buchanan Street, Albany, CA 94710, U.S.A.
R.N. Chibbar. Plant Biotechnology Institute, National Research Council Canada, 110 Gymnasium Place, Saskatoon, SK S7N OW9, Canada.

[1] Corresponding author (e-mail: ygu@pw.usda.gov).

doi: 10.1139/G03-071

de la séquence nucléotidique du contig. De plus, la D hordéine de l'orge a été comparée aux HMW-gluténines du blé au niveau de la conservation des résidus cystéine et de l'organisation des domaines répétés.

Mots clés : évolution, HMW-gluténines, structure du génome, rétrotransposon.

[Traduit par la Rédaction]

Introduction

The prolamin storage proteins of barley, wheat, and rye account for about half of the total grain protein. In barley (*Hordeum vulgare*), the major prolamin fraction is composed of hordeins, which are further divided into several heterogenous groups, namely D (high molecular weight), C (sulphur-poor), B, and γ (sulphur-rich) hordein peptides. Nevertheless, all hordein proteins share certain biochemical and physical characteristics. For example, they are insoluble in water or dilute salt solutions, soluble in alcohol–water mixtures, and have high glutamine and proline content. These properties are determined in part by the presence of similar repeat domains in the prolamin polypeptides. Numerous genes for prolamin proteins have been isolated from a variety of cereals, including those encoding B, C, D, and γ hordein. From a detailed sequence analysis, it was proposed that sulphur-rich and sulphur-poor hordeins are derived from the same ancestor gene (Shewry et al. 1995). In barley, as with prolamin protein genes in other cereals, transcripts for B, C, γ, and D hordein genes were shown to be coordinately expressed in the developing starchy endosperm tissue (Sørensen et al. 1989). Analysis of the promoter regions for prolamine protein genes has revealed extensive sequence similarity, which may contribute to their similar transcript expression patterns (Sørensen et al. 1989), further suggesting close evolutionary relationships among prolamin protein genes. Moreover, barley prolamin protein genes are present as multiple copies in the genome, with the exception of the *Hor*3 locus encoding the D hordein gene (Shewry et al. 1985; Kanazin et al. 1993). Genetically, hordein genes all map to chromosome 5(1H), with the D hordein gene on the long arm and other hordein protein genes on the short arm. However, little is known about the molecular mechanism underlying the structure and evolution of regions containing genes for prolamin proteins in the barley genome.

Barley and wheat genomes are closely related with respect to their evolutionary origin. While barley (*H. vulgare*) has a true diploid genome, bread wheat (*Triticum aestivum*) is a hexaploid species consisting of three homoeologous genomes (termed A, B, and D), with each genome having seven mostly colinear chromosomes. Comparative studies using RFLP markers have revealed that the linkage groups of wheat and barley genomes are remarkably conserved (Dubcovsky et al. 1996). The chromosomal regions carrying the hordein genes on chromosome 5(1H) are likely orthologous to the regions on the group 1 chromosome of each wheat genome. The corresponding region in the wheat genomes contains similar seed storage protein genes to those found in barley. For instance, on the long arms of group 1 chromosomes reside genes encoding the wheat high molecular weight (HMW) glutenin subunits that are homologous to D hordein. The S-rich γ gliadins and low molecular weight (LWM) glutenins (evolutionarily part of the gliadin superfamily), and S-poor ω gliadins located on the short arms of group 1 chromosomes are the counterparts of the B, C, and γ hordeins.

Analysis of genomic sequence can provide a detailed view of the composition, organization, and evolution of plant genomes (Bennetzen 2000). As a step toward understanding the molecular basis of Triticeae genome evolution, particularly in the regions containing prolamin storage-protein genes, we previously reported the sequence of a BAC clone of 102 kb containing the HMW-glutenin locus from a diploid wheat, *Ae. tauschii*, the D-genome donor of hexaploid wheat (Anderson et al. 2003). Here, we report the sequence of a barley BAC clone of 120 kb containing the D hordein locus. The annotated sequence is compared with its orthologous region from the D genome of *Ae. tauschii.*

Materials and methods

BAC clone isolation and sequencing

Barley EST clone HVSMi0015C21f (GenBank accession No. BG367985), which encodes a D hordein, was used to screen a set of 17 high-density filters printed with barley BAC library clones constructed from the modern cultivar *Hordeum vulgare* 'Morex' (Yu et al. 2000). Hybridization was performed with the ^{32}P-labeled D-hordein probe at 65 °C in a solution containing 0.5 mol sodium phosphate/L (pH 7.2), 7% w/v SDS, 1% w/v bovine serum albumin (BSA), and 1 mmol EDTA/L. After hybridization for 20 h, the filters were washed three times in 0.5× SCC for 30 min/wash before exposure to X-ray films. Positive clones and their corresponding library addresses were identified using the Hybsweeper program developed in house (G.R. Lazo, N. Liu, Y.Q. Gu, and O.D. Anderson, unpublished). A total of 12 positive clones were recovered from the screening. Southern blot and fingerprinting analyses on these BAC clones were performed to select a desirable clone that would provide maximum sequence information flanking the D hordein locus. Clone 184G9 was chosen because of its large size and central location in the assembled contig (data not shown).

A shotgun subcloning and subsequent sequencing approach was used to determine the sequence of the selected BAC clone. The BAC DNA was first isolated using the large construct kit procedure (Qiagen, Valencia, Calif.), in which an exonuclease treatment is included to remove *Escherichia coli* genomic DNA contamination. The purified BAC DNA was sheared into two different sizes with an average of 3 and 8 kb, respectively. The sheared fragments were blunt ended with mung bean nuclease (BioLab, Beverly, Mass.) and dephosphorylated with shrimp alkaline phosphatase (USB, Cleveland, Ohio). Single A tails were added by incubating with *Taq* polymerase in the presence of dNTPs. DNA inserts were ligated into pCR4TOPO vectors using the TA cloning kit (Invitrogen, Carlsbad, Calif.). The resulting DNA was electroporated into DH10B electroMAX cells (Invitrogen). Single colonies were picked and grown in deep-well microtiter blocks containing 1.5 mL YT medium (8 g of bacto-tryptone, 5 g of bacto-yeast extract, and 2.5 g of NaCl in 1 L) with 50 mg carbenicillin/L in a HIGro shaker (Genemachines, Menlo Park, Calif.) at 37 °C with shaking speed set at 4.5 for 20 h. The plasmid DNA was isolated using the PerfectPrep Direct Bind Kit (Eppendorf, Boulder, Colo.). Inserts were sequenced from both directions with T7 and T3 primers using BigDye terminator chemistry (Applied Biosystems, Foster City, Calif.) and run on an ABI Prism 3700 capillary sequencer. After the initial assembling of the sequences, there were three gaps, all of which had G- and (or) C-rich borders. Gaps were filled by primer walking using dGTP BigDye terminator chemistry (Applied Biosystems).

Sequence analysis

The sequences were assembled using the Lasergene SeqMan module (DNAStar, Madison, Wis.) (www.DNAStar.com). In this module, we set the stringency for base calling and quality assessment to “high” to generate the most accurate consensus sequence possible. The quality sequences were assembled using a 40-bp window size and a 96% match requirement. We had previously assembled a 102-kb contiguous sequence containing the HMW-glutenin locus of the wheat D genome using the Lasergene SeqMan module (Anderson et al. 2003). To confirm the sequence assembly for the barley D-hordein BAC clone, a restriction digest of the BAC DNA was performed using *Hin*dIII, *Eco*RI, and *Not*I. The digest patterns were then compared with predicted restriction patterns of the computer-assembled sequence.

For annotation, the finished sequence was compared with NCBI nonredundant and dbEST databases using BLASTn, BLASTx, and tBLASTx algorithms, version 2.2.4 (Altschul et al. 1997) to search for additional genes. In addition, FGENESH (http://www.softberry.com/berry.phtml) and GENESCAN (http://genes.mit.edu/Genscan.html) were used for gene prediction. Mobile elements were identified by multiple BLAST searches as described above. In most cases, known mobile elements in wheat and barley can be found by comparison against the Triticeae repeat sequence database (TREP) at the GrainGenes Web site (http://wheat.pw.usda.gov/ITMI/repeats/).

Results and discussion

BAC sequencing and assembly

The original assembly of the region around the barley D-hordein locus was accomplished by sequencing both ends of randomly sheared subclones. A total of 6004 sequences resulted in 4 contigs with 3 gaps. The failure to join these contigs together was due to the presence of highly G- and C-rich regions that cause sequencing reaction termination. Primers were designed to sequence through these gaps with dGTP BigDye terminator chemistry. In all cases, the gaps were filled. The final assembled sequence for the barley BAC clone is 120 652 bp long with an average coverage of 21 times the genome (21×). We also assembled the BAC clone at two other levels of sequence coverage (10× and 15×) to evaluate the best balance between coverage and gap number. At 10× coverage, five gaps were obtained; at 15× coverage, the gap number was the same as that at 21× coverage. The final contiguous sequence is deposited in GenBank under accession No. AY268139.

Gene content and distribution

A combination of gene prediction by GENESCAN, FGENESH, and BLAST searches was used to search for putative genes in the 120-kb contiguous sequence. A putative gene is defined if the sequence shows homology with the characterized genes and both gene-finding programs predict a complete gene construct. BLAST searches for matching ESTs were performed to support the prediction of the genes. Four putative genes were identified in the sequenced BAC clone as illustrated in Fig. 1. The gene encoding the D hordein storage protein spans between position 62 032 and 64 305. Like other prolamin protein genes, the D hordein gene possesses no introns. The translated peptide contains 578 amino acids and shows a high sequence similarity with the wheat HMW-glutenins (see later discussion). A BLAST search against the barley EST database revealed 619 perfect matches to the sequence of the D-hordein gene (data not shown). These ESTs are all derived from the barley endosperm or endosperm-associated tissue libraries, suggesting tissue specificity of the D hordein promoter. The promoter of the D hordein gene has been previously characterized. The results indicated that the high and tissue-specific expression of the D hordein gene was not associated with the methylation status of genomic promoter sequence (Sørensen et al. 1996). Recently, it has been shown that transgenes driven by the D-hordein promoter were more stably inherited than transgenes driven by the rice actin promoter when producing transgenic barley plants (Cho et al. 2002). Such a promoter could be a useful resource for the improvement of barley grains by means of transgene technology (Zhong 2001).

A second storage protein gene encoding a seed globulin was found to reside at position 38 912, approximately 23 kb from the D hordein gene. The DNA separating these two storage protein genes is a cluster of retrotransposable elements (Fig. 1 and see later discussion). A BLAST search against the EST database identified 10 EST clones (data not shown). These ESTs were assembled into a single contig with a sequence identical to the full-length globulin sequence. A seed-globulin-encoding gene has not yet been characterized in barley. BLAST searches showed a strong similarity to the previously reported *Ae. tauschii* seed globulin (score 269, 2×10^{-68}) (Anderson et al. 2003). As in the wheat globulin gene, there are no introns in the barley globulin gene. The translated peptide with 224 amino acids also showed matches to a rice α-globulin (X63990; 2×10^{-9}) (Shorrosh et al. 1992) and a maize α-globulin (AF 371278; 7×10^{-14}) recently discovered in a maize EST sequencing project (Woo et al. 2001). As seen in Fig. 2, the amino acid sequence alignment reveals high similarity among cereal globulins. In particular, there are eight cysteine residues that were found to be conserved in all the sequences. Cysteines are known to be conserved residues in prolamin storage proteins, because they are essential for the formation of

Fig. 1. Arrangements of genes and mobile elements at the barley D-hordein locus. Yellow boxes represent putative genes, the arrow in the box indicates the transcription orientation. An uninterrupted retroelement is diagrammed as a colored box flanked by two arrows of the same color. The direction of the arrow indicates the transcription orientation of the retroelement. The nomenclature for new transposable elements was according to the methods described by SanMiguel et al.(2002). In the names of retroelements, S stands for solo LTR when the target signature duplication sites were identified. P refers to a partial retroelement with deletions of portions of the LTR and (or) coding region.

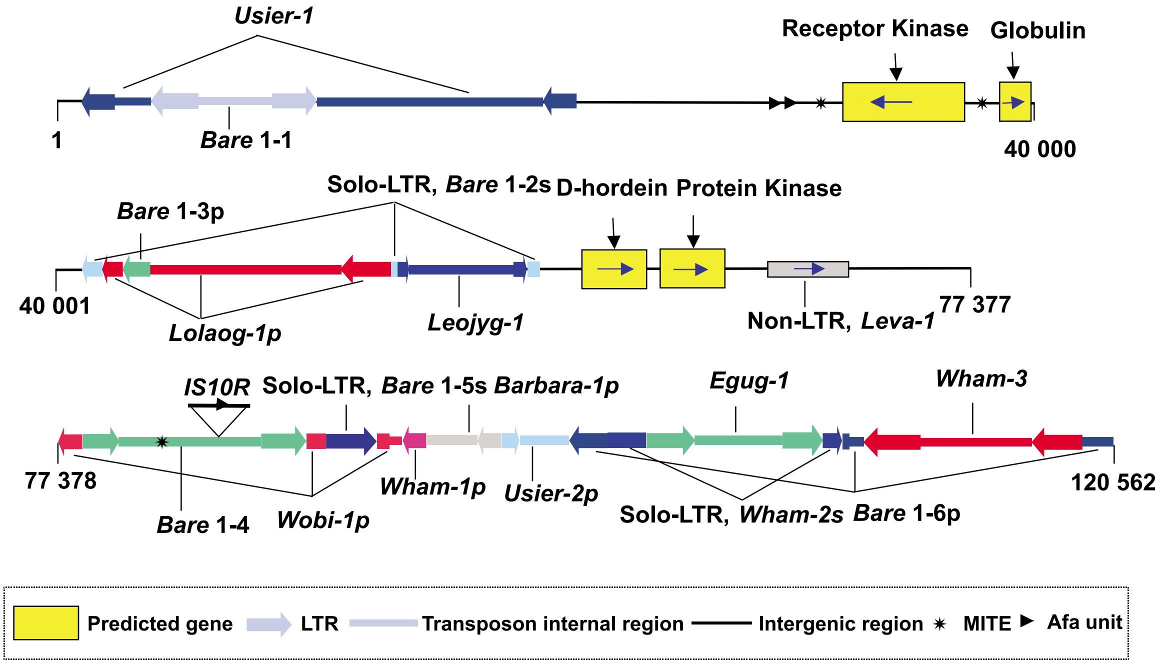


Fig. 2. Comparison of globulin sequences among barley, wheat, maize, and rice. The alignments were done using ClustalW analysis. Cysteine residues conserved in all the four globulin sequences are marked with stars. Conserved amino acids are shaded black. In the case that amino acids conserved in barley and wheat differ from those conserved in rice and maize, only the amino acids in wheat and barley are shaded.

```
  1  MG-KFIFFAVFL--TTLVTISAAQG-VLEQSLADAQCRGEVQAKPLLACRQILEHQLTGRA----VGVRP
  1  MG-RFVFFAVFL--AALVAGSAAQG-VLEQSLTDAQCRGEVQEKPLLACRQILEQQLTGRAGEGAVGVPL
  1  MAKIAAAAAAALCFAALVAVAVCQGEVERQRLRDLQCWQEVQESPLDACRQVLDRQLTGGGG--GGGVGP
  1  MASKVVFFAAAL-MAAMVAISGAQLSESEMRFRDRQCQREVQDSPLDACRQVLDRQLTGRER--FQPMFR
                                         ★           ★

 63  FQAQWGARDRCCQQLESVSRGCRCSALRGMVRDYEQSMP-PLREGRRRSSGERQQE-QGCSGEST-AEQQ
 67  FHAQWGARERCCRQLESVSRECRCAALRGMVRDYEQSMP-PLGEGRHGSSGERQQE-QGCSGESTEPEQR
 69  FRWGTGLRMRCCQQLQDVSRECRCAAIRSMVRGYEEAMP-PLEKGWWPWGRQQQPPPQGGGGG-------
 68  RPGALGLRMQCCQQLQDVSRECRCAAIRRMVRSYEESMPMPLEQGWSSSSSEYYGG-EGSSSE-------
               ★★         ★ ★

130  QEVQGGQYGSETGESQQQQGGGYHGVTVGRGGQQQGQMLCRERPQRQQQGEGFSGEGAQ-QKPKVGRVRL
135  QEVQGGQYGSETG-GGQQQGGGYHGVTVGRGGQRQGQVLCHKRPQRQQ-GEGFSGEGAQ-QKPQAGRVRL
131  ---QGGYYYPCSR-----PGEGYG---YGQGGQRQMYPPC--RP-------GTTGGG-----PRIGRVRL
130  ---QG--YYGEGS-----SEEGY----YGEQQQ----------P-------GMT------------RVRL

199  TKVR-----LPTACRI-EPQECSVFSTLPVLG  barley
202  TKVR-----LPTACRI-EPQECSVFTADQYY   wheat
176  TKAREYAAGLPMMCRLSEPQECSIFSGGDQY   maize
158  TRARQYAAQLPSMCRV-EPQQCSIFAAG-QY   rice
                  ★       ★
```

intramolecular and intermolecular disulfide bonds and have been used to categorize prolamin proteins into S-rich and S-poor classes (Shewry and Tatham 1990). The high conservation of amino acid sequences, particularly these cysteine resides, suggests that they are the orthologous genes of different cereals. On the other hand, sequence divergence can be observed in the central regions where both barley and wheat globulins have extended sequences rich in glutamine residues as compared with maize and rice. The increased length of the central regions in wheat and barley may be due to a series of amplifications of the nucleotide sequences encoding glutamines. In addition, the amino acid alignment of

Fig. 3. Sequence alignment of barley leucine-rich repeat receptor kinase with rice Xa21. The alignment was done using ClustalW analysis. Asterisks indicate identical residues, colons indicate conserved amino acid substitutions, and dots indicate semi-conserved substitutions. The first amino acid of each imperfect leucine-rich repeat based on analysis of Xa21 (Song et al. 1995) is marked with solid triangles. The serine (threonine) kinase region is boxed.

the sequences suggests that the rice globulin is more related to maize than to wheat or barley globulins (Fig. 2).

A third gene recognized in the sequence encodes a putative receptor protein kinase. Both GENESCAN and FGENESH predicted the same gene with two introns spanning from position 27 835 to 31 616. The translated peptide shows a high similarity to several receptor protein kinase sequences (*E* value = 0). Among them is the well-characterized Xa21 protein, the product of a rice resistance (R) gene conferring gene-for-gene resistance to the bacterial pathogen *Xanthomonas oryzae* pv. *oryzae* (Song et al. 1995). Xa21 belongs to the class of plant leucine-rich repeat receptor kinases. It has been well documented that this class of receptor kinases has a diverse array of biological functions, ranging from plant development to responses to biotic and abiotic stresses (Song et al. 1995; Jinn et al. 2000; Montoya et al. 2002; Asai et al. 2002). The leucine-rich receptor kinases are composed of several structural domains including leucine-rich repeat domains, transmembrane domains, and a cytoplasmic protein kinase domain. Pairwise comparison with Xa21 showed that the barley receptor kinase protein contains similar leucine-rich repeats and conserved kinase domains (Fig. 3). The biological function of this barley leucine-rich receptor kinase is unknown. However, two EST matches (CA015492 and BQ466381) were identified through BLAST searches. Both ESTs are present in the barley endosperm library from imbibed seeds, suggesting a potential role for the gene in endosperm development.

The fourth gene, only 500 bp downstream from the D hordein gene, encodes a putative serine (threonine) protein kinase. Both GENESCAN and FGENESH predicted the same peptide with a size of 583 amino acids. The product shows a high similarity to several putative protein kinases in the database. The best match is to a rice putative serine (threonine) protein kinase with an expect value of 1.0×10^{-141} (BAB89770). BLAST searches against the barley EST database did not recover sequences exactly matching the exon regions of the putative kinase gene. However, less significant EST matches were identified (e.g., BQ754294, 3×10^{-13}), which in turn encode putative protein kinases (data not shown). It is possible that the mRNA encoding this protein kinase is rare and (or) only expressed under certain conditions or in certain tissues, and that the current EST sequencing efforts have not yet recovered such transcripts or that it is not expressed. The comparison of the barley serine (threonine) protein kinase with the orthologous gene of *Ae. tauschii* revealed that they share 96% amino acid sequence identity (data not shown). Such a high conservation implies that both genes are active (Anderson et al. 2003), although further studies will be required to prove that these protein kinases represent functional genes.

The overall gene density in D-hordein locus region is one gene per 30-kb based on the presence of four genes in a 120-kb sequence. These four genes are located in the central region of the contig sequence (Fig. 1). The two gene-containing regions or gene islands are separated by a 23-kb region composed mainly of a cluster of retrotransposons (see later discussion). However, in the colinear region of the *Ae. tauschii* D genome, the active globulin gene is immediately followed by the Dy HMW-glutenin gene (Anderson et al. 2003). The presence of retroelement insertions between the globulin and D hordein genes in the barley genome suggests that the size and gene density of gene islands are subject to change during evolution. The observed gene density calculated on the basis of several BAC sequences of Triticeae genomes is thus far approximately 8–42 kb/gene (Wicker et al. 2001; Wei et al. 2002). Two BAC sequences representing regions of high gene density were both derived from chromosome segments containing disease resistance loci, in which genes were clustered as a result of frequent R gene duplication (Wei et al. 2002; Brooks et al. 2002). Sequence regions with less gene density have been usually found to contain clusters of retrotransposons inserted in the intergenic regions (Wicker et al. 2001; Ramakrishna et al. 2002)

Repetitive DNA in intergenic regions

Sequence analysis of BAC clones containing large segments of genomic DNA from several cereal species have begun to delineate the composition, organization, and evolution of large, complex cereal genomes (Shirasu et al. 2000; Wicker et al. 2001; Dubcovsky et al. 2001; Wei et al. 2002; Ramakrishna et al. 2002; Anderson et al. 2003). It is generally agreed that transposable elements make up most of the intergenic regions and that the genome size is directly correlated with the number of retroelements in the plant genome. We identified a total of 16 LTR retrotransposon elements in the D-hordein BAC contig sequence (see Fig. 1; Table 1). They account for 63% of the contig sequence. The Ty1-*copia*-type *Bare*-1 retrotransposon is the most dominant LTR element in the region with five *Bare*-1 copies, two of which are full-length (*Bare*-1-1 and *Bare*-1-4). These five *Bare*-1 elements account for 20% of the total contig sequence. This high percentage of sequence composition may reflect *Bare*-1 elements being the major, active retrotransposable elements in the barley genome as they constitute 2.8% of the barley genomic sequence and are widely distributed throughout the genome (Vicient et al. 1999). In the D hordein locus, the much higher than average percentage of *Bare*-1 elements may indicate local activity of *Bare*-1 retrotransposons. Such local activity has been observed in other regions of the barley and wheat genomes (Wei et al. 2002; Wicker et al. 2001).

The second most prevalent class of retroelements in this region is the Ty3-*gypsy*-like *Wham* class retrotransposon. The *Wham* retroelement was recently discovered in the diploid wheat genome (A^m), and its activity and functionality in plant genomes has not been well characterized (SanMiguel et al. 2002). We identified three *Wham* retroelements in the

```
Barley  METRPWLLRLIAILTTTTALLLHPSTSSS-VSTAHDLPALLSFKSLITKDPLGALSSWTT 59
        * : * ** :: :     **** **:*..  .:* *  ******* :  :   :*:**.*
Rice    MISLPLLLFVLLFS----ALLLCPSSSDDDGDAAGDELALLSFKSSLLYQGGQSLASWNT 56

Barley  NGSTHGFCSWTGVECSS---AHPGHVKALRLQGLGLSGTISPFLGNLSRLRALDLSGNKL 116
        .*    .*:*.** *.     ** :*  * *:. .*** *** ***** ** ***..* *
Rice    SG-HGQHCTWVGVVCGRRRRRHPHRVVKLLLRSSNLSGIISPSLGNLSFLRELDLGDNYL 115
                                    ▲      ▲       ▲        ▲      ▲

Barley  QGQIPSSIGNCFALRTLNLSVNSLSGAIPPAMGNLSKLLVLSVSKNDISGTIPTSFAG-L 175
        .*:**..:..   *: *:** **:.*:**.*:*  :**  *.:*:*:: * ** .:.. *
Rice    SGEIPPELSRLSRLQLLELSDNSIQGSIPAAIGACTKLTSLDLSHNQLRGMIPREIGASL 175
               ▲        ▲      ▲       ▲        ▲      ▲        ▲

Barley  ATVAVFSVARNHVHGQVPPWLGNLTALEDLNMADNIMSGHVPPALSKLINLRSLTVAINN 235
        :: : : :* : *::*. *****:*:::::: * :** :*.:*.:* .* ::.:. **
Rice    KHLSNLYLYKNGLSGEIPSALGNLTSLQEFDLSFNRLSGAIPSSLGQLSSLLTMNLGQNN 235
             ▲      ▲       ▲        ▲      ▲       ▲        ▲

Barley  LQGLIPPVLFNMSSLECLNFGSNQLSGSLPQDIGSMLPNLKKFSVFYNRFEGQIPASLSN 295
        *.*:**  ::*:***..:.. .*:*.* :* :  . *  *: :.:  ***.*:****::*
Rice    LSGMIPNSIWNLSSLRAFSVRENKLGGMIPTNAFKTLHLLEVIDMGTNRFHGKIPASVAN 295
        ▲       ▲        ▲      ▲        ▲        ▲      ▲       ▲

Barley  ISSLEHLSLHGNRFRGRIPSNIGQSGRLTVFEVGNNELQATESRDWDFLTSLANCSSLLL 355
         * *  :.::** * * *.*.:*:  .** : : .* :*: *. **.*::.*:***.*
Rice    ASHLTVIQIYGNLFSGIITSGFGRLRNLTELYLWRNLFQTREQDDWGFISDLTNCSKLQT 355
               ▲     ▲       ▲     ▲         ▲             ▲       ▲

Barley  VNLQLNNLSGILPNSIGNLSQKLEGLRVGGNQIAGLIPTGIGRYLKLAILEFADNRFTGT 415
        :**  ***.*:****:.*** .*. * :  *:*:* **..**. : *  * :.:*.* *:
Rice    LNLGENNLGGVLPNSFSNLSTSLSFLALELNKITGSIPKDIGNLIGLQHLYLCNNNFRGS 415
               ▲        ▲        ▲      ▲       ▲        ▲      ▲

Barley  IPSDIGKLSNLKELSLFQNRYYGEIPSSIGNLSQLNLLALSTNNLEGSIPATFGNLTELI 475
        :**.:*:*.**  *  ::*.  *.** :****::**:* *.**::.* ** *:.***:*:
Rice    LPSSLGRLKNLGILLAYENNLSGSIPLAIGNLTELNILLLGTNKFSGWIPYTLSNLTNLL 475
            ▲        ▲      ▲       ▲        ▲      ▲       ▲

Barley  SLDLASNLLSGKIPEEVMRISSLALFLNLSNNLLDGPISPHIGQLANLAIIDFSSNKLSG 535
        **.*::* *** **.*::.*.:*::::*:*:* *:*.*. .**:* **. :. .**:***
Rice    SLGLSTNNLSGPIPSELFNIQTLSIMINVSKNNLEGSIPQEIGHLKNLVEFHAESNRLSG 535
         ▲      ▲       ▲         ▲      ▲       ▲        ▲      ▲

Barley  PIPNALGSCIALQFLHLQGNLLQGQIPKELMALRGLEELDLSNNNLSGPVPEFLESFQLL 595
         ***:**.*  *::*:**.***.*.**. *  *:*** ****.***** :*  * .: :*
Rice    KIPNTLGDCQLLRYLYLQNNLLSGSIPSALGQLKGLETLDLSSNNLSGQIPTSLADITML 595
             ▲        ▲      ▲       ▲        ▲      ▲       ▲

Barley  KNLNLSFNHLSGPVPDKGIFSNASVISLTSNGMLCGGPVFFHFPTCPYPSPDKLASHKLL 655
        :.****** : * **  * *: ** **: .*. ****   :*:* *  *  ::   . :*
Rice    HSLNLSFNSFVGEVPTIGAFAAASGISIQGNAKLCGGIPDLHLPRC-CPLLENRKHFPVL 654
          ▲      ▲      ▲        ▲

Barley  QILVFTAVGAFILLGVCIAARCYVNKSRGDAHQDQENIPEMFQRISYTELHSATDSFSEE 715
         * *  *..  ** .: :    : ...:* . : . :    .  :**::* .***.*:
Rice    PISVSLAAALAILSSLYLLITWHKRTKKGAPSRTSMKG---HPLVSYSQLVKATDGFAPT 711

Barley  NLVGRGSFGSVYKGTSGSGANLITAAVKVLDVQRQGATRSFISECNALKMIRHRKLVKVI 775
        **:* *********. .       .*****.::.  * :** :**:**: :***:***::
Rice    NLLGSGSFGSVYKGKLNIQD---HVAVKVLKLENPKALKSFTAECEALRNMRHRNLVKIV 768

Barley  TVCDSLDHSGNQFKALVLEFIPNGSLDKWLHPSTEDEFGT--PNLMQRLNIALDVAEALE 833
        *:*.*:*: **:***:* :*:*****:.*:**.*:*: .    ** :*:.* **** **:
Rice    TICSSIDNRGNDFKAIVYDFMPNGSLEDWIHPETNDQADQRHLNLHRRVTILLDVACALD 828

Barley  YLHDHIDPPIVHCDVKPSNILLDDDMVAHLGDFGLAKIIRAEKSKQSLADQSCSVGIKGT 893
        *** *   *:****:*.**:***.*****:******:*:   .*   : :.:.*:*: **
Rice    YLHRHGPEPVVHCDIKSSNVLLDSDMVAHVGDFGLARILVDGTS--LIQQSTSSMGFIGT 886

Barley  IGYVAPEYGTGTEISVEGDVYSYGVLLLEMLTGRRPTDPFFSDTTNLPKYVEMACPGNLL 953
        ***.*****.*   *..**:****:*:**::**:****. *    .* :***:.  *.:
Rice    IGYAAPEYGVGLIASTHGDIYSYGILVLEIVTGKRPTDSTFRPDLGLRQYVELGLHGRVT 946

Barley  ETMDVNIRCNQE---------PQAVLELFAAPVSRLGLACCRGSARQRIKMGDVVKELGA 1004
        :.:*.::  :.*         *   :    . : ****:*.:  . .*   **::.**.*
Rice    DVVDTKLILDSENWLNSTNNSPCRRITECIVWLLRLGLSCSQELPSSRTPTGDIIDELNA 1006

Barley  IKQIIMASQNYASWSTKLY 1023
        *** : .    .. .:  :
Rice    IKQNLSGLFPVCEGGSLEF 1025
```

Table 1. Retrotransposons and the chronology of their insertions in the D-hordein region.

Elements	Type	Length (bp)	5′ LTR[a]	3′ LTR	Total base substitution[b]	Time (million years)[c]
Bare-1-1	*copia*	8 608	1 817	1 816	27	1.2
Bare-1-2S	*copia*	1 778				
Bare-1-3P	*copia*	766				
Bare-1-4	*copia*	8 717	1 826	1 837	63	2.6
Bare-1-5S	*copia*	1 726				
Bare-1-6P	*copia*	3 293				
Usier-1	*copia*	11 968	1 349	1 418	51	3.1
Usier-2P	*copia*	3 513				
Barbara-1P	*copia*	2 194				
Leojyg-1	*copia*	4 598	141	141	2	1.0
Wham-1P	*gypsy*	638				
Wham-2S	*gypsy*	1 376				
Wham-3	*gypsy*	9 802	1 581	1 587	74	3.5
Lolaog-1P	*gypsy*	1 078				
Egug-1	unclassified	7 292	1 519	1 533	68	3.5
Wobi-1P	unclassified	3 421				

[a] Full length retrotransposons with two LTRs were used to calculate the time of insertion.

[b] Deletion of insertion was counted as a single event of base substitution.

[c] Divergence time is calculated as $K/2 \times K_{\mathrm{nus}}$.(Wei et al. 2002). K_{nus} is the base substitution rate per nucleotide per year. A value of $K_{\mathrm{nus}} = 6.5 \times 10^{9}$substitutions/site/year derived from the grass (Gramineae) *adh*1–*adh*2 region (Guat et al. 1996) was employed in the calculation.

contig sequence: one full length, and the other two interrupted or partial. The full-length *Wham*-3 element was inserted in the coding region of the *Bare*-1-6p element. All three *Wham* elements are located at the 3′ end of the contig sequence, where they are arranged in complex nested clusters (Fig. 1). This nested retrotransposon organization forms a pattern similar to that previously reported for retrotransposons in maize and Triticeae genomes (SanMiguel et al. 1996; Wicker et al. 2001; SanMiguel et al. 2002; Anderson et al. 2003). The nested retrotransposon structure reported here spans a 43-kb region from the *Wobi*-1p element to the end of the contig sequence (Fig. 1).

We discovered and named three new LTR retroelements. *Usier* is a *copia*-like retrotransposon with an LTR size of about 1.4 kb. There are two copies of *Usier* elements present in our BAC clone. The first one, located at the 5′ end of the contig, was interrupted by the insertion of a *Bare*-1-1 element. The second, a partial *Usier-2p* element with one full-length LTR, was identified in the nested retroelement region at the end of the contig. *Leojyg-1* is a *copia*-like retrotransposon with a size of 4598 bp. The small size is attributed to the short LTRs with sizes of 141 bp, among the smallest we are aware of for LTR retrotransposable elements. The full-length *Leojyg-1* element was used for a BLASTn search and identified a previously unidentified partial *Leojyg* element in one barley BAC sequence deposited in the NCBI database (AF47071), with 91% nucleotide identity over a 2.5-kb region. The third new element is *Lolaog-1p*, which is a partial *copia*-like retroelement resulting from retroelement insertion and deletion events (see Fig. 1). It appears that a *Bare*-1 element was inserted into the 5′ LTR region of the *Lolaog* element and then a deletion event removed most of the *Bare*-1 sequence and part of the 5′ LTR of *Lolaog-1p*. We were able to define the full-length 3′ LTR based on the position of a 5-bp target site duplication (TSD) immediately flanking both LTRs.

Although retrotransposable elements with LTRs are the most abundant in large cereal genomes, non-LTR retrotransposons have been discovered as ubiquitous components of many plant genomes (Schmidt 1999). Because they lack distinct long terminal repeats, the borders of non-LTR retrotransposon insertion sites are not easily defined. Sequence analysis revealed that there is one non-LTR retrotransposon, *Leva-1*, in the region downstream of the serine (threonine) protein kinase gene. Its borders were recognized by comparison with the corresponding region of the *Ae. tauschii* D genome sequence. There is a 3-bp TSD found at the borders of the element. At one of its ends, there is a stretch of A residues preceding the TSD, a typical structure for non-LTR retroelements (Vershinin et al. 2002).

Miniature inverted repeat transposable elements (MITEs) are mobile elements that are distinguished by their small size and close association with plant genes (Bureau and Wessler 1994). At the end of the receptor protein kinase gene, there is one MITE of the *Stowaway* family that is followed by two *Afa* unit tandem repeats. A second MITE was inserted in the region between the receptor kinase and globulin. A third MITE was inserted in the *Bare*-1-4 retroelement.

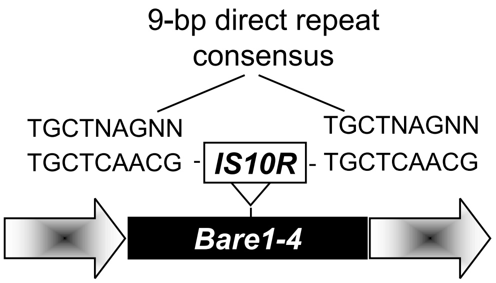
In the nested retroelement region at the 5′ end of the contig, we identified *Wobi-1p* and *Egug-1* elements, which do not seem to belong to any recognized class of retrotransposons. *Wobi-1p* was identified by comparing its sequence to new elements discovered in a BAC sequence from a tetraploid wheat BAC sequence (Y. Gu, O. Anderson, X. Kong, and D. Coleman-Derr, unpublished data). *Egug-1* is a full-length element, which has two LTRs with sizes of 1.5 kb, flanked by perfect 5-bp TSDs. BLASTx searches revealed that what would normally be the coding region of these elements does not encode any known peptides. This type of element has been reported previously and may represent a new group of mobile elements in the plant genome (Wicker et al. 2001).

Plant genomes vary tremendously in size. Triticeae genomes, including barley (5000 Mb), have large genome sizes compared with 130 Mb in *Arabidopsis* and 450 Mb in rice (*Oryza sativa*). The presence of large numbers of retroelements in the sequenced region supports the notion that retrotransposon amplification contributes to genome expansion (SanMiguel et al. 1996; SanMiguel et al. 1998). However, plants must have mechanisms that counterbalance the genome expansion caused by retroelement amplification. One such mechanism is the unequal crossing-over and (or) intrachromosomal recombination between LTRs, which can delete most of the sequence of the targeted retroelement (Shirasu et al. 2000; Devos et al. 2002). One of the products resulting from this type of recombination is the solo LTR. In the barley 66-kb *Rar* region, most of the retroelements exist as solo LTRs (Shirasu et al. 2000). However, in other Triticeae genome regions, solo LTRs are not common (SanMiguel et al. 2002). We found three solo LTRs, as confirmed by their 5-bp TSDs, *(Bare*-1-2s, *Bare*-1-5s, and *Wham-2s*), in different regions. Other types of deletions with unknown mechanisms must exist, which also contribute to the removal of retroelement sequences. In our BAC clone, one deletion occurred between *Lolaog-1p* and *Bare*-1–3p in the nested retroelement region in the middle of the contig sequence. Another deletion event was found between the *Wham-1p* and *Barbara-1p*, leaving a small portion of *Wham* 3′-LTR and a partial *Barbara* element (Fig. 1). In addition, small deletions, insertions, and base-pair substitutions are frequently observed in the coding regions of these elements. These mutations caused translation frame shifts or premature stop codons. Therefore, none of the LTR retroelements identified in the contig sequence are likely to be functional components. Disruption of a retroelement's activity can stop its continual replication and transposition and then prevent its rapid amplification in the host genome. Other mechanisms such as methylation in repetitive sequences or gene silencing can also assist plants in resisting retroelement amplification (Hirochika et al. 2000).

Insertion of *IS10R* into the BAC sequence

Initial BLAST searches against the NCBI database revealed that a segment of the BAC sequence spanning from position 84 049 to position 85 378 had strong similarity to a sequence that is present in many eukaryotic clones. Further analysis showed that it matched exactly to the 1329-bp *IS10R* sequence. *IS10R* is the right side of the flanking inverted repeat sequence of bacterial transposon *Tn10*, which has a total length of 9147 bp. The mobile element of *Tn10* is *IS10R*, which possesses the gene for transposase, thus enabling the movement of *IS10R* as an independent entity. The distribution of *Tn10–IS10R* in prokaryotic genomes and plasmids has been well studied and the *IS10R* elements were found to be more frequent than the intact *Tn10* transposon (Bender and Kleckner 1992). However, there are no reports of the presence of *Tn10*/*IS10R* sequences in eukaryotic genomes. It is likely that this *IS10R* sequence was inserted into the BAC clone during the BAC library construction and subsequent manipulation. The addition of the 1329-bp *IS10R* appeared to have no effect on the contig assembling of the 12 positive BAC clones. Recently, it has been reported that a considerable portion of the DNA databases are contaminated with *IS*-derived sequences (Bender and Kleckner 1992; Kovaøåk et al. 2001). Full-length *IS10R* insertions are often flanked by 9-bp direct repeats with a preferred consensus sequence of 5′-TGCTNAGNN-3′. A 9-bp direct repeat with a sequence of 5′-TGCTCAACG-3′ was found to be present at both ends of the intact *IS10R* element, which was inserted in the *Bare*-1-4 retrotransposon element in our sequence (see Figs. 1 and 4). The 9-bp repeats allow for easy removal of the *IS10R* sequence from the barley BAC sequence data. The 120 562-bp sequence in the NCBI database is the corrected version for the barley D hordein BAC locus, and no longer contains the *IS10R* sequence.

Fig. 4. Insertion of *IS10R* in the D hordein BAC clone. *IS10R* was inserted in the coding region of the *Bare*-1–4 retroelement. The 9-bp direct repeats flanking the *IS10R* element was indicated. The 9-bp preferred consensus for the *IS10R* insertion site was provided based on the results from Kovaøåk et al. (2001).

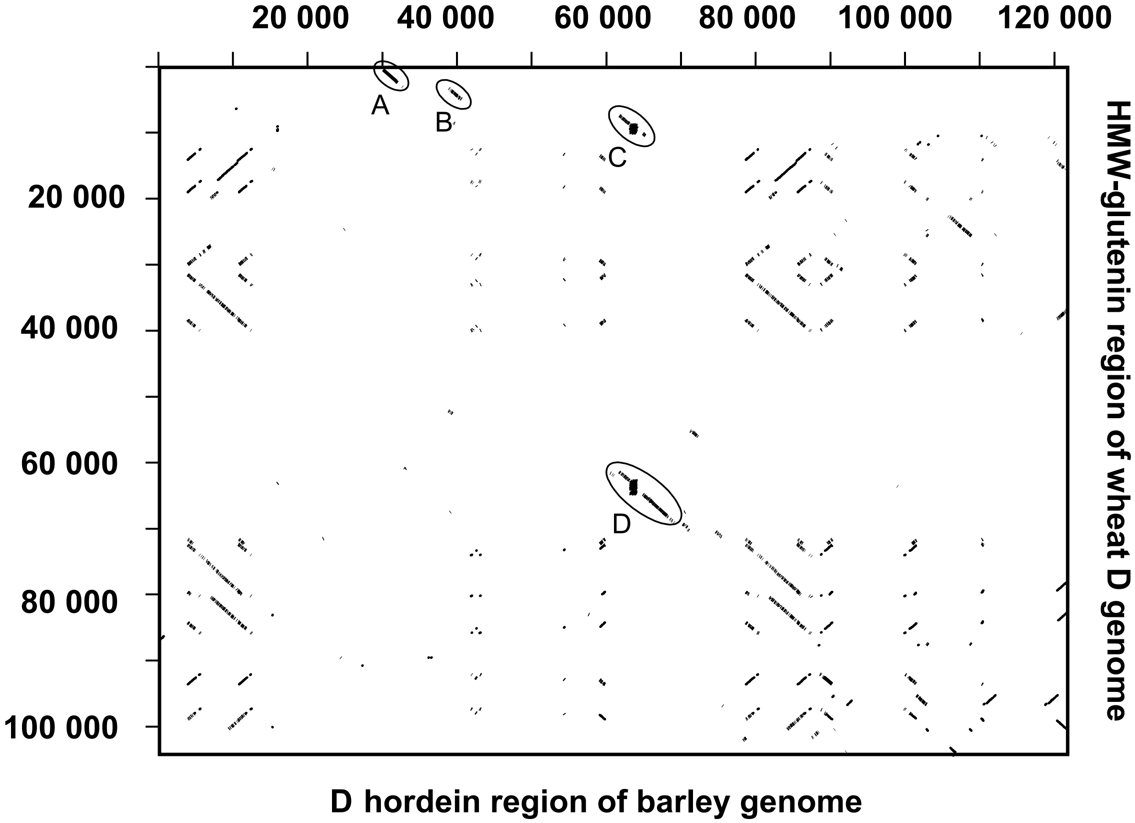


It has been documented that an insertion of *Tn10–IS10R* can cause rearrangement of the acceptor genome, including deletions, inversions, and other mutations (Bogosian et al. 1993; Chalmers and Kleckner 1996). A BLAST search revealed that the *Bare*-1-4 element had a strong match to the *Bare*-1_AF427791-d retroelement found in the *Mla* locus in the barley genome (Wei et al. 2002). Sequences flanking the *IS10R* insertion site were carefully compared with the sequence of *Bare*-1_AF427791-d. No evidence of sequence arrangements was detected in the *Bare*-1-4 region containing the *IS10R* sequence.

Comparison with the orthologous region of the wheat D genome

The wheat and barley genomes are closely related in origin and evolution. They represent two major cereal crops of the Triticeae grass tribe. The chromosomal regions containing prolamin storage protein genes are of particular importance because of their contributions to grain quality. Recently, we reported the sequence analysis of the wheat D genome HMW-glutenin locus of *Ae. tauschii*, a diploid wheat ancestor and the D-genome donor of hexaploid wheat (Anderson et al. 2003). A direct comparison of the genomic region containing the D-hordein locus with its orthologous region of the wheat D genome will allow for the understanding of the genome's evolution, particularly in the storage protein regions. A dotplot analysis was performed using the barley se-

Fig. 5. Dotplot of barley D hordein and *Ae. tauschii* HMW-glutenin BAC clone sequences. Sequence match criteria were 60% over a 50-bp window. Colinear sequence regions are circled and marked with letters A, B, C, and D.



quence against its orthologous region of the *Ae. tauschii* D genome (Fig. 5). Conserved sequences were restricted to four regions. Region A contains the leucine-rich repeat receptor kinase and region B contains the globulin gene. Regions C and D represent the intersection of the D-hordein sequence with each of the paralogous Dy and Dx HMW-glutenin genes of the D genome. Region D also contains the putative serine (threonine) protein kinase. The gap in region D is due to the presence of a central nonrepetitive region, followed by the second repetitive domain unique to barley D hordein (see later discussion). In the compared regions, the gene content and order are conserved, with the exception of HMW-glutenin duplication in the wheat D genome. Highly related sequences were also found in several non-colinear regions. These regions are comprised of sequences of highly related LTR retrotransposable elements, but characterized with different names in wheat and barley. For example, the sequence of the *Bare*-1 element in barley is highly similar to that of the *Angela* element in wheat (Wicker et al. 2001).

Sequence comparison reveals that although the barley D-hordein region contains orthologous genes as in the HMW-glutenin locus or the wheat D genome, sequences in the intergenic regions are not colinear (Fig. 5). To gain a better understanding of dynamic sequence changes, we compared the sequences of barley and wheat genomes in the regions between the receptor kinase and the serine (threonine) protein kinase genes (Fig. 6). A detailed examination revealed that in the D-genome sequence, a region containing the globulin and HMW-glutenin genes was duplicated and inserted in an adjacent region, an event that lead to the presence of paralogous Dy and Dx HMW-glutenin genes in wheat genomes (Anderson et al. 2003). This important gene duplication could be responsible for conferring to wheat the physical and biochemical properties that make it suitable for a variety of food applications, including breadmaking. The duplication probably occurred after the divergence of wheat and barley from their common ancestor, since all wheat genomes are known to have both paralogous HMW-glutenin genes. It is less likely that the duplication event occurred before speciation and that a single deletion event removed the original duplicated sequence in the barley genome. The presence of both the x- and y-type HMW-glutenin orthologs in rye suggests that the gene duplication occurred before the separation of the rye and wheat genomes (De Bustos et al. 2001). This is consistent with evolutionary studies that have shown the wheat genomes to be more closely related to rye than to barley (Huang et al. 2002).

A second noticeable difference between the compared regions is that in barley, there are 35.6 kb between the start of the receptor kinase and the end of the protein kinase; in the wheat D genome, that distance is 64.7 kb. Gene duplication is one cause of size expansion. However, the original duplicated region containing the globulin and HMW-glutenin genes was about 7.2 kb in size (Y. Gu, O. Anderson, X. Kong, and D. Coleman-Derr, unpublished). Comparative sequence analysis reveals that the remainder of the size difference was mainly due to the activity of transposable elements. For example, in barley, there is only one nested

Fig. 6. Sequence organizations of (A) the barley D-hordein locus and (B) its orthologous region of the *Ae. tauschii* D genome. The solid boxes represent putative genes with arrows indicating transcription orientation. The solid box bar marked with an asterisk is the pseudogene of the second globulin in the HMW-glutenin region of the wheat D genome. The horizontal solid back lines are the intergenic regions into which retroelements have been inserted. The vertical solid black lines indicate the position of retroelement insertions.

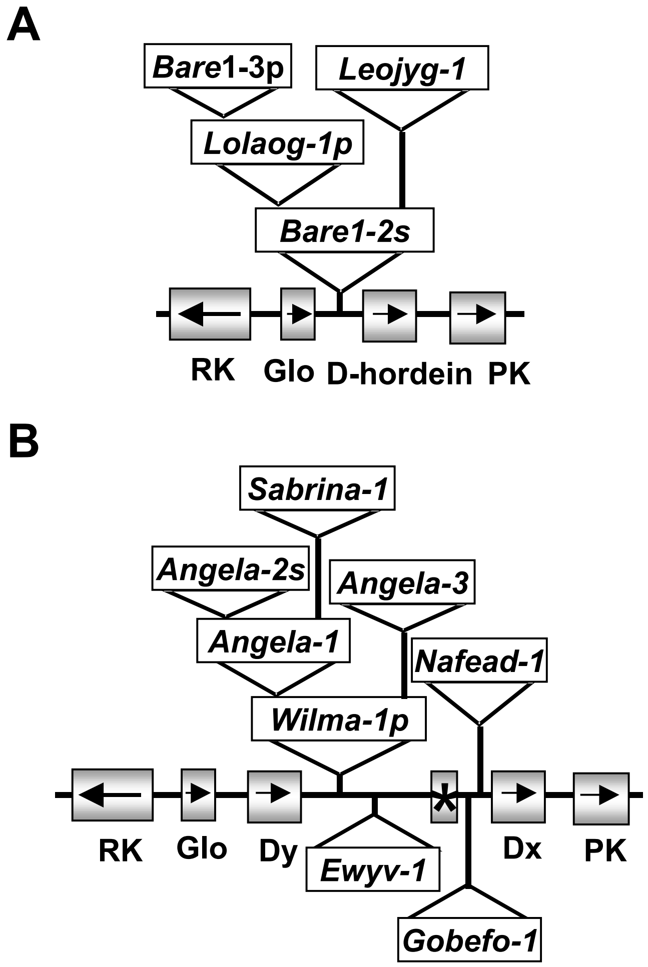
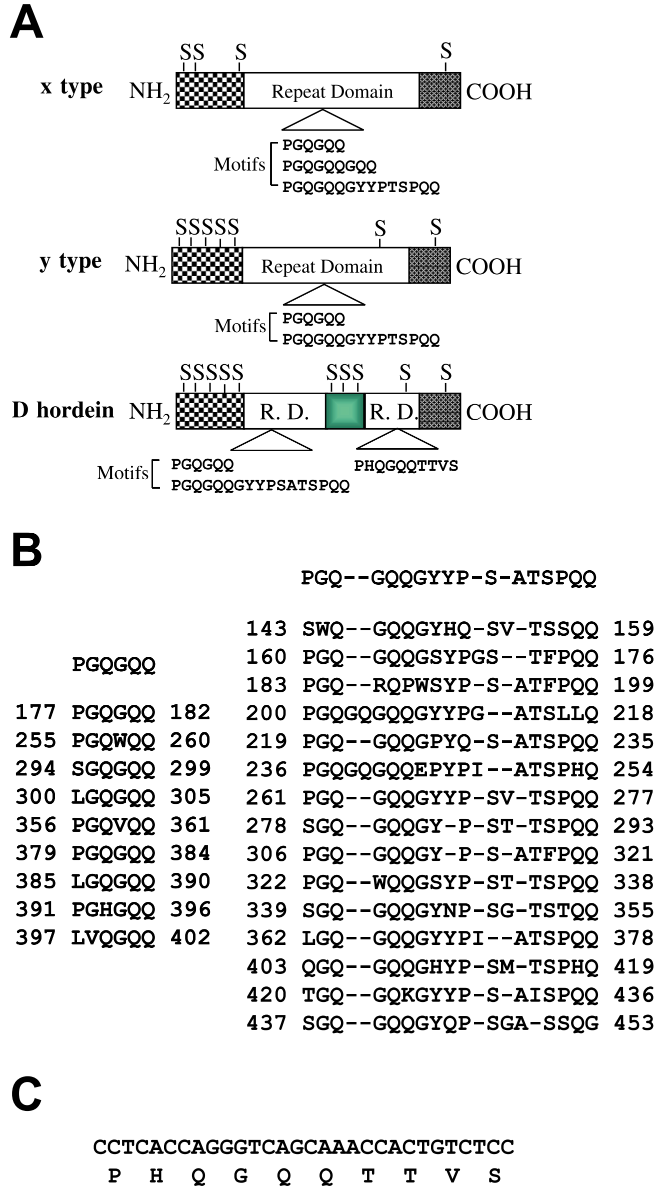


Fig. 7. Comparison of the repeat domains of D hordein to those of wheat HMW glutenins. (A) Structural domains for y- and x-type HMW glutenins and D hordein. (B) Repeat sequence organization of the first repeat domain in D hordein. The consensus for the repeat sequences in the first repeat domain is aligned and positions of amino acid resides are indicated. (C) The nucleotide sequence repeat and the encoded amino acid sequence repeat of the second repeat domain in D hordein.



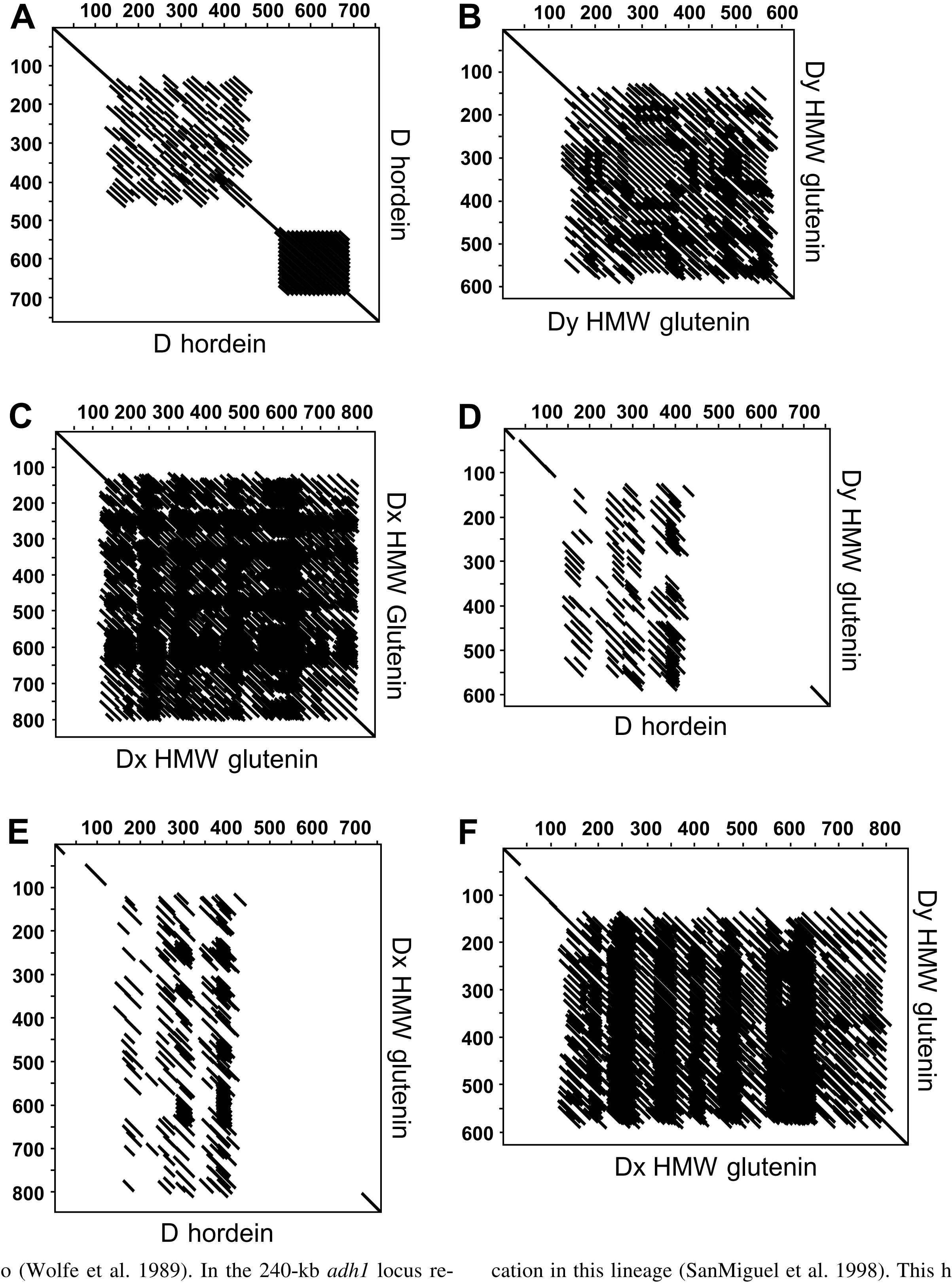
```
                      PGQ--GQQGYYP-S-ATSPQQ

                  143 SWQ--GQQGYHQ-SV-TSSQQ 159
    PGQGQQ        160 PGQ--GQQGSYPGS--TFPQQ 176
                  183 PGQ--RQPWSYP-S-ATFPQQ 199
177 PGQGQQ 182    200 PGQGQGQQGYYPG--ATSLLQ 218
255 PGQWQQ 260    219 PGQ--GQQGPYQ-S-ATSPQQ 235
294 SGQGQQ 299    236 PGQGQGQQEPYPI--ATSPHQ 254
300 LGQGQQ 305    261 PGQ--GQQGYYP-SV-TSPQQ 277
356 PGQVQQ 361    278 SGQ--GQQGY-P-ST-TSPQQ 293
379 PGQGQQ 384    306 PGQ--GQQGY-P-S-ATFPQQ 321
385 LGQGQQ 390    322 PGQ--WQQGSYP-ST-TSPQQ 338
391 PGHGQQ 396    339 SGQ--GQQGYNP-SG-TSTQQ 355
397 LVQGQQ 402    362 LGQ--GQQGYYPI--ATSPQQ 378
                  403 QGQ--GQQGHYP-SM-TSPHQ 419
                  420 TGQ--GQKGYYP-S-AISPQQ 436
                  437 SGQ--GQQGYQP-SGA-SSQG 453
```

```
CCTCACCAGGGTCAGCAAACCACTGTCTCC
 P  H  Q  G  Q  Q  T  T  V  S
```

retroelement region spanning 16.9-kb between the globulin and D-hordein genes (Fig. 6A), whereas in the wheat D genome, the two paralogous Dy and Dx genes are 51.9 kb apart, and the separating DNA includes a 31-kb cluster of retroelements. Additionally, there are three non-LTR retroelements inserted into the region between the globulin and D-hordein genes, one immediately following the nested retroelement cluster, the other two between the globulin pseudogene and the Dx HMW-glutenin gene (Fig. 6B; Anderson et al. 2003).

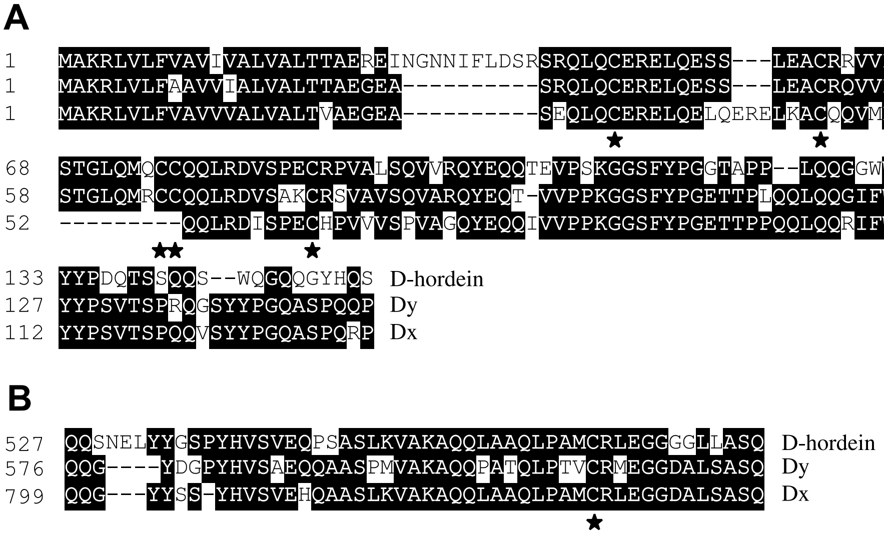
When the retroelements present in the sequenced regions were compared, none of the retroelements present in barley were colinear with the sequences in the orthologous region of the wheat D genome, suggesting that all retroelements were inserted after speciation. The insertion time of an LTR retroelement can be estimated based on the assumption that two LTRs of the retroelement were identical at the time of insertion and that differences between them directly reflect the time of evolution. There are five elements that possess both the intact 5′ and 3′ LTRs in the sequenced barley region. They were used to estimate insertion times as shown in Table 1. *Wham-3* and *Egug-1* were inserted 3.5 million years ago. The youngest retroelement is *Leojyg-1*. It was inserted into the *Bare*-1-2s element approximately one million years ago. This is consistent with the estimation that wheat and barley diverged from each other between 10 and 14 million

Fig. 8. Dot matrix analysis of D hordein and the Dx and Dy HMW-glutenin proteins. Sequences of D hordein and the Dx and Dy HMW glutenins were used to perform pairwise comparison against themselves and each other as shown in each plot (A, B, C, D, E, and F). Sequence match criteria was 50% over a 20-amino-acid window.



years ago (Wolfe et al. 1989). In the 240-kb *adh1* locus region of the maize genome, the insertion times of 14 out of the 16 retroelements that can be dated are all within the last three million years, suggesting recent retroelement amplification in this lineage (SanMiguel et al. 1998). This hypothesis is supported by the evidence presented here that the retrotransposon elements are not colinear in the barley and *Ae. tauschii* D genomes and that the insertion times of all of

Fig. 9. Sequence alignments of the non-repetative regions of barley D hordein and *Ae. tauschii* Dx and Dy HMW-glutenin proteins. The alignments were done using ClustalW analysis. Sequences of the N- and C-terminal regions were used for comparison. (A) Alignment of the N-terminal region sequences. (B) Alignment of the C-terminal region sequences. Stars represent conserved cysteine residues. Shaded boxes represent conserved amino acid sequences.


A
1 MAKRLVLFVAVIVALVALTTAEREINGNNIFLDSRSRQLQCERELQESS---LEACRRVVDQQLVGQLPW
1 MAKRLVLFAAVVIALVALTTAEGEA-----------SRQLQCERELQESS---LEACRQVVDQQLAGRLPW
1 MAKRLVLFVAVVVALVALTVAEGEA-----------SEQLQCERELQELQERELKACQQVMD---------
68 STGLQMQCCQQLRDVSPECRPVALSQVVRQYEQQTEVPSKGGSFYPGGTAPP--LQQGGWWGTS---VKW
58 STGLQMRCCQQLRDVSAKCRSVAVSQVARQYEQT-VVPPKGGSFYPGETTPLQQLQQGIFWGTSSQTVQG
52 ---------QQLRDISPECHPVVVSPVAGQYEQQIVVPPKGGSFYPGETTPPQQLQQRIFWGIP-ALLKR
133 YYPDQTSSQQS--WQGQQGYHQS D-hordein
127 YYPSVTSPRQGSYYPGQASPQQP Dy
112 YYPSVTSPQQVSYYPGQASPQRP Dx
B
527 QQSNELYYGSPYHVSVEQPSASLKVAKAQQLAAQLPAMCRLEGGGGLLASQ D-hordein
576 QQG----YDGPYHVSAEQQAASPMVAKAQQPATQLPTVCRMEGGDALSASQ Dy
799 QQG---YYSS-YHVSVEHQAASLKVAKAQQLAAQLPAMCRLEGGDALSASQ Dx

the retroelements that can be dated are within the last 3.5 million years.

Comparison of D hordein with x- and y-type HMW-glutenin subunits

The wheat HMW-glutenin genes are the most studied wheat genes because they encode critical proteins that directly contribute to the physical and chemical properties of wheat doughs (Shewry et al. 1992). The x- and y-type HMW-glutenin subunits are highly similar, since they diverged from a common progenitor after their duplication. They have a similar three-domain structure (Fig. 7A). The short N- and C-terminal domains are nonrepetitive in their amino acid sequence, while the central domain is composed of repeats of simple peptide motifs. The repeat motifs are variants of 6, 9, and 15 amino acids (Fig. 7A).

D hordein, the major storage protein in barley endosperm, is homologous to the wheat HMW glutenins. The D-hordein gene encodes a peptide with 758 amino acid residues. Dotplot analysis of the D-hordein peptide against itself is shown in Fig. 8A. The block-like structures indicate where repetitive motifs were found to have multiple matches within compared sequences. Two block-like structures are present in the plot, indicating that D hordein has two distinct repeat regions separated by one non-repeat region (Fig. 8A). In contrast, neither the Dy nor the Dx HMW-glutenin has a central non-repeat region (Fig. 8B and 8C). While the sequences that constitute the central non-repeat region and the second repeat domain found in the D-hordein sequence are not conserved in the wheat HMW glutenins (Fig. 8D and 8E), the sequence of the first repeat domain in D hordein is similar to the repetitive domain of HMW glutenin. However, the Dx and Dy HMW glutenins have similar repeat sequences (Fig. 8F). Furthermore, the sequences at both N- and C-terminal regions of these prolamin storage proteins are conserved as indicated by the dotplot analyses (Figs. 8D–8F).

To more closely examine the evolution of these prolamin proteins, the amino acid sequences of the N- and C-terminal regions of D hordein and the Dy and Dx HMW glutenins were aligned (Figs. 9A and 9B). High sequence similarities were detected among the compared sequences with few exceptions. One region containing two cysteines in the Dx HMW glutenin was deleted. Barley D hordein contains this region, leading us to conclude that the deletion event occurred after the duplication of HMW glutenin. In prolamin proteins, the conserved cysteine residues are believed to have a direct impact on dough quality. Previously, it was reported that the Dx HMW-glutenin subunits of most cultivars possess three cysteine residues in the N-terminal region instead of the five present in the Dy subunit (Fig. 7A ; Shewry et al. 1992). Our comparative analysis on these orthologous and paralogous sequences provided molecular evidence to support the idea that the differences arose during evolution. Apart from these two cysteine residues in the deleted region, the three cysteines at the N termini and one at the C termini are highly conserved. However, both D hordein and Dy HMW glutenin have an additional, non-colinear, cysteine in the 3′ nonrepetitive region. These cysteine residues were likely derived from different mutation events.

As suggested by Fig. 8, the central nonrepetitive region is unique to D-hordein. Furthermore, sequence analysis revealed the presence of three cysteine residues in this region (Fig. 7A). The effect of these unique cysteine residues on the functionality of D-hordein is unclear. It is known that barley flour will form weak dough when mixed with water. These additional cysteine residues in the central region could have a negative impact on dough strength, since they may facilitate the formation of intramolecular disulfide bonds with the cysteine residues at both N- and C-terminal regions. It can be predicted that intrachain disulfide bonds will prohibit formation of intermolecular disulfide bonds, resulting in short and weak polymers, whereas the interchain disulfide bonds allow increased polymer formation and additional branching.

The repeat sequences in the first repetitive regions of D hordein are aligned in Fig. 7B. The first repetitive region is composed of repeat motifs that are similar to those found in wheat HMW glutenins. One hexapeptide motif, PGQGQQ, appears nine times in the first repetitive region. In three of the nine repeats, the first-position proline was changed to leucine, which was caused by a point mutation of the second nucleotide in the codon from C to T. The second repeat motif is made up of the hexapeptide motif plus a relatively less conserved sequence containing 11–16 amino acids (Fig. 7B). It is apparent that variations in the middle of the second repeat motif were caused by several insertions or deletions of amino acids. An examination of the nucleotide sequences demonstrated that these amino acid misalignments were caused by insertions and deletions (indels) in multiples of three nucleotides (data not shown). These indel triplets have maintained a full-length open reading frame during evolution. Such indels were observed in the repetitive regions of the wheat HMW-glutenins (Anderson and Greene 1989).

Surprisingly, in the second repetitive region of D hordein, the 10-amino-acid repeat motif (PHQGQQTTVS) is perfectly repeated 14 times both at nucleotide and amino acid levels, indicating a relatively recent evolutionary origin or some type of homogenizing mechanism (Fig. 7C). Conformational analysis by secondary structure prediction and by circular dichroism spectroscopy of synthetic peptides suggested that the TTVS repeats in D hordein form β pleated sheets (Halford et al. 1992). However, the question regarding the impact of the non-repeat region between the two repetitive domains on the conformation of overall protein structure still needs to be addressed.

Despite the fact that dotplot analysis showed differences between the two repetitive regions in the D-hordein sequence (Fig. 8), a tetrapeptide sequence, QGQQ, is highly conserved among all the repeat motifs. Similar short motif sequences are often recognized as repeats and contribute to the high content of glutamines in prolamin storage proteins (Shewry et al. 1995). The biological function of these glutamine-rich motifs is not well understood. Considering that nutrients for seedling growth during germination are provided by the storage deposits in the endosperm, including these prolamin proteins, it is expected that some structural features of the storage proteins allow for easy mobilization of energy reserves. The cysteine endoproteases are among the most abundant proteases in the germinating barley endosperm. The prolamin storage proteins, including D hordein, are likely the natural substrates of these proteases. The cleavage specificity for the cysteine endoproteases EP-A and EP-B has been characterized in barley. Several cleavage sites similar to QGQQ have been predicted for D hordein (Davy et al. 2000). A detailed understanding of the structure and diversity of prolamin storage proteins is an important prerequisite for attempts to manipulate grain quality, because it indicates the extent to which the structure of the proteins can be modified without affecting their biological properties.

Acknowledgements

We are grateful to Dr. Fusheng Wei for the generous help in annotation of the sequence and to Dr. Frank You for the assistance in BAC contig assembly. We also thank Devin Coleman-Derr and Curt Crossman for their critical reading of the manuscript.

References

Altschul, S.F., Maden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, M., and Lipman, D.J. 1997. Gapped BLAST and PSI-PLAST: a new generation of protein database search program. Nucleic Acids Res. **25**: 3389–3402.

Anderson, O.D., and Greene, F.C. 1989. The characterization and comparative analysis of high MW glutenin genes from genomes A and B of hexaploid wheat. Theor. Appl. Genet. **77**: 689–700.

Anderson, O.D., Rausch, C., Moullet, O., and Lagudah, E.S. 2003. The wheat D-genome HMW-glutenin locus: BAC sequencing, gene distribution, and retrotransposon cluster. Funct. Integr. Genomics, **3**: 56–68.

Asai, T., Tena, G., Plotnikova, J., Willmann, M.R., Chiu, W.L., Gomez-Gomez, L., Boller, T., Ausubel, F.M., and Sheen, J. 2002. MAP kinase signalling cascade in *Arabidopsis* innate immunity. Nature (London), **415**: 977–983.

Bender, J., and Kleckner, N. 1992. Tn10 insertion specificity is strongly dependent upon sequences immediately adjacent to the target-site consensus sequence. Pro. Natl. Acad. Sci. U.S.A. **89**: 7996–8000.

Bennetzen, J. 2000. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell, **12**: 1021–1029.

Bogosian, G., Bilyeu, K., and O'Neil, J.P. 1993, Genome rearrangement by residue *IS10* elements in strains of *Escherichia coli* K-12 which have undergone *Tn10* mutagenesis and fusaric acid selection. Mol. Gen. Genet. **233**: 17–22.

Brooks, S.A., Huang, L., Gill, B.S., and Fellers, J.P. 2002. Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance. Genome, **45**: 963–972.

Bureau, T.E., and Wessler, S.R. 1994. Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. U.S.A. **91**: 1411–1415.

Chalmers, R., and Kleckner, N. 1996. IS10/Tn10 transposition efficiently accommodates diverse transposon end configuration. EMBO J. **15**: 5112–5122.

Cho, M.-J., Choi, H.-W., Jiang, W., Ha, C.D., and Lemaux, P.G. 2002. Endosperm-specific expression of green flurescent protein driven by the hordein promoter is stably inherited in transgenic barley (*Hordeum vulgare*) plants. Physiol. Plant, **115**: 144–154.

Davy, A., Sorensen, M.B., Svendsen, I., Cameron-Mills, V., and Simpson, D. 2000. Prediction of protein cleavage sites by the barley cysteine endoproteases EP-A and EP-B based on the kinetics of synthetic peptide hydrolysis. Plant Physiol. **122**: 137–145.

De Bustos, A., Rubio, P., and Jouve, N. 2001. Characterization of two gene subunits on the 1R chromosome of rye as orthologs of each of the *Glu-1* genes of hexaploid wheat. Theor. Appl. Genet. **103**: 733–742.

Devos, K.M., Brown, J.K.M., and Bennetzen, J.L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. **12**: 1075–1079.

Dubcovsky, J., Luo, M.C., Zhong, G.Y., Bransteitter, R., Desai, A., Kilian, A., Kleinhofs, A., and Dvorak, J. 1996. Genetic map of diploid wheat, *Triticum monococcum* L., and its comparison with maps of *Hordeum vulgare* L. Genetics, **143**: 983–99.

Dubcovsky, J., Ramarkrishna, W., SanMiguel, P., Busso, C.S., Yan, L., Shiloff, B.A., and Bennetzen, J.L. 2001. Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. Plant Physiol. **125**: 1342–1353.

Guat, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. 1996. Substitution rate comparisons between grasses and palm: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc. Natl. Acad. Sci. U.S.A. **93**: 10274–10279.

Halford, N.G., Tatham, A.S., Sui, E., Daroda, L., Dreyer, T., and Shewry, P.R. 1992. Identificationof a novel β-turn-rich repeat motif in the D-hordeins of barley. Biochim. Biophys. Acta, **1122**: 118–122.

Hirochika, H., Miura, H., and Sawada, S. 2000. Silencing of retrotransposons in *Arabidopsis* and reactivation by the *ddm1* mutation. Plant Cell, **12**: 357–368.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. 2002. Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. Proc. Natl. Acad. Sci. U.S.A. **99**: 8133–8138.

Jinn, T.L., Stone, J.M., and Walker, J.C. 2000. HAESA, an *Arabidopsis* leucine-rich repeat receptor kinase, controls floral organ abscission. Genes Dev. **14**: 108–117.

Kanazin, V., Ananiev, E., and Blake, T. 1993. Variability among members of the *Hor-2* multigene family. Genome, **36**: 397–403.

Kovaøåk, A., Matzke, M.A., Matzke, A.J.M., and Koukalová, B. 2001. Transposition of *IS*10 from the host *Escherichia coli* genome to a plasmid may lead to cloning artifacts. Mol. Gen. Genomics, **266**: 216–222.

Montoya, T., Nomura, T., Farrar, K., Kaneta, T., Yokota, T., Bishop, G.J. 2002. Cloning the tomato *curl3* gene highlights the putative dual role of the leucine-rich repeat receptor kinase tBRI1/SR160 in plant steroid hormone and peptide hormone signaling. Plant Cell, **14**: 3163–3176.

Ramakrishna, W., Dubcovsky, J., Park, Y.J., Busso, C., Emberton, J., SanMiguel, P., and Bennetzen, J.L. 2002. Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. Genetics, **162**: 1389–1400.

SanMiguel, P., Tiknonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake Berhanm, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., and Bennetzen, J.L. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science (Washington, D.C.), **274**: 765–768.

SanMiguel, P., Gaut, B.S., Thhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. Nat. Genet. **20**: 43–45.

SanMiguel, P.J., Ramakrishna, W., Bennetzen, J.L., Busso, C.S., and Dubcovsky, J. 2002. Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). Funct. Integr. Genomics, **2**: 70–80.

Schmidt, T. 1999. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant. Mol. Biol. **40**: 903–910.

Shewry, P.R., and Tatham, A.S. 1990. The prolamin storage proteins of cereal seeds: structure and evolution. Biochem. J. **267**: 1–12.

Shewry, P.R., Bunce, N.A., Kreis, M., and Forde, B.G. 1985. Polymorphism at the *Hor1* locus of barley (*Horeum vulgare*) Biochem. Genet. **23**: 391–404.

Shewry, P.R., Halford, N.G., and Tatham, A.S. 1992. High molecular weight subunits of wheat glutenin. J. Cereal Sci. **15**: 105–120.

Shewry, P.R., Napier, J.A., and Tatham, A. 1995. Seed storage proteins: structure and biosynthesis. Plant Cell, **7**: 945–956.

Shirasu, K., Schulman, A.H., Lahaye, T., and Schulze-Lefert, P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res. **10**: 908–915.

Shorrosh, B.S., Wen, L., Huang, J.L., Pan, J.S., and Hermodson, M.A. 1992. A novel cereal storage protein: molecular genetics of the 19-kDa globulin of rice. Plant Mol. Biol. **18**: 151–154.

Song, W.-Y., Wang, G.-L., Chen, L.-L., Kim, H.-S., Pi, L.-Y., Holstern, T., Gardner, J., Wang, B., Zhai, W.-X., Zhu, L.-H., Fauquet, C., and Ronald, P. 1995. A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. Science (Washington, D.C), **270**: 1804–1806.

Sørensen, M.B., Cameron-Mills, V., and Brandt, A. 1989. Transcriptional and post-translational regulation of gene expression in developing barley endosperm. Mol. Gen. Genet. **217**: 195–201.

Sørensen, M.B., Muller, M., Skerritt, J., and Simpson, D. 1996. Hordein promoter methylation and transcriptional activity in wild-type and mutant barley endosperm. Mol. Gen. Genet. **250**: 750–760.

Vershinin, A.V., Druka, A., Alkhimova, A.G., Kleinhofs, A., and Heslop-Harrison, J.S. 2002. LINEs and *gypsy*-like retrotransposons in *Hordeum* species. Plant Mol. Biol. **49**: 1–14.

Vicient, C.M., Kalendar, R., Anamthawat-Jonsson, K., Suoniemi, A., and Schulman, A.H. 1999. Structure, functionality, and evolution of the *Bare*-1 retrotransposon of barley. Genetica, **107**: 53–63.

Wei, Fusheng, Wing, R.A., and Wise, R.P. 2002. Genome dynamics and evolution of the *Mla* (powdery mildew) resistance locus in barley. Plant Cell, **14**: 1903–1917.

Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E., and Keller, B. 2001. Analysis of a contiguous 211-kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J. **26**: 307–316.

Wolfe, K.H., Gouy, M., Yang, Y.W., Sharp, P.M., and Li, W.H. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. U.S.A. **86**: 6201–6205.

Woo, Y.M., Hu, D.W., Larkin, B.A., and Jung, R. 2001. Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. Plant Cell, **12**: 2297–2317.

Yu, Y., Tomkins, J.P., Waugh, R., Frisch, D.A., Kudrna, D., Kleinhofs, A., Brueggeman, R.S., Muehlbauer, G.J., and Wing, R.A. 2000. A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. Theor. Appl. Genet. **101**: 1093–1099.

Zhong, G.-Y. 2001. Genetic issues and pitfalls in transgenic plant breeding. Euphytica, **118**: 137–144.